**SRI SIDDHARTHA ACADEMY OF HIGHER EDUCATION**

***(Declared as Deemed to be University Under Section 3 of the UGC Act, 1956* Approved by AICTE, Accredited by NBA, NAAC ‘A’ Grade*)***

**AGALKOTE, TUMAKURU – 572107 KARNATAKA**



## Project Report On

## **“PREDICTION OF COVID CASES IN MAHARASHTRA STATE USING LINEAR REGRESSION”**

**Submitted by**

**SHREYAS T M (18CS088)**

**VIVEK V R (18CS117)**

**THEJASWI K (18CS104)**

**VINOD S C (18CS114)**

**Under the guidance of**

**CHANNABASAVARAJU T P**

**Assistant Professor**

**Dept. of Computer Science & Engineering**

****S.S.I.T, Tumakuru.

**DEPARTMENT OF COMPUTER SCIENCE**

SRI SIDDHARTHA INSTITUTE OF TECHNOLOGY

**(A Constituent College of SRI SIDDHARTHA ACADEMY OF HIGHER EDUCATION)**

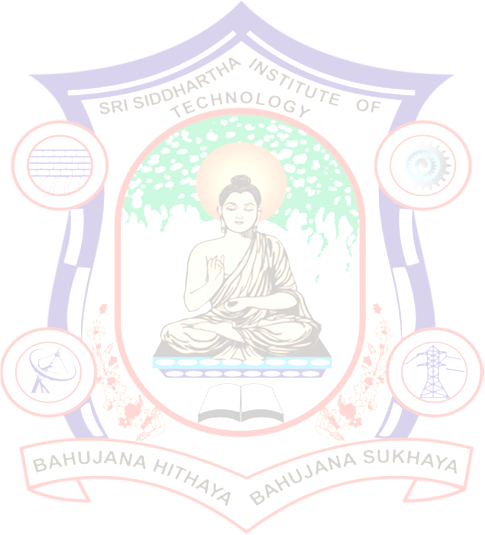
### MARALUR, TUMAKURU -572105

### 2020-21

**SIDDHARTHA INSTITUTE OF TECHNOLOGY**

**(A Constituent College of Sri Siddhartha Academy of Higher Education)**

**MARALUR, TUMAKURU-572105**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

CERTIFICATE

This is to certify that the project entitled**“PREDICTION OF COVID CASES IN MAHARASHTRA STATE USING LINEAR REGRESSION”** is a bonafide work carried out by**SHREYAS T M , VIVEK V R , THEJASWI K** *and* **VINOD S C** in the partial fulfillment for the award of thedegree of Bachelor of Engineering in *“****Computer Science and Engineering”*** during the academic year of 2020-2021.

It is certified that all of the corrections/suggestions indicated for the internal assessments have been incorporated in the report. The project report has been approved as it satisfies the academic requirements with respect to the project work prescribed for the award of the degree of Bachelor of Engineering in **Computer Science and Engineering.**

|  |  |  |
| --- | --- | --- |
| Signature of Guide | Signature of HOD | Signature of Principal |
| **Channabasavaraju T P** | **Dr. M. Siddappa** | **Dr. M.S. Raviprakasha** |
| B.E., M.Tech., | B.E., M.Tech., Ph.D., MISTE | B.E., M.Tech., Ph.D., |
| Asst. Professor, | Professor & Head, | Principal, |
| Dept. of CSE, SSIT, Tumakuru. | Dept. of CSE, SSIT, Tumakuru. | SSIT, Tumakuru |

**External Examiners:- Signature with date:-**

**1.**

**2.** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**ACKNOWLEDGEMENT**

At the outset we express our most sincere grateful thanks to holy sanctum of **“SRI SIDDHARTHA INSTITUTE OF TECHNOLOGY”**, the temple of learning, for giving us an opportunity to pursue the degree course in Computer Science and Engineering thus help shaping our career.

We wish to express my sincere gratitude to **Dr. M. S. RAVIPRAKASHA** Principal, Sri Siddhartha Institute of Technology, for providing us an excellent academic environment, which has nurtured our practical skills, and for kindly obliging our requests and providing timely assistance.

We are highly grateful to **Dr. M. SIDDAPPA,** Dean academics and Head of Department of Computer Science and Engineering, for patronizing us throughout the project and for his encouragement and moral support.

We also wish to express our deep sense of gratitude to our project guide **CHANNABASAVARAJU T P,** Assistant Professor, Department of Computer Science and Engineering, for his valuable suggestions, guidance, moral support and encouragement in completion of this project successfully. We have been fortunate for having his precious help.

Finally, we express our gratitude to all the teaching and non-teaching of the **Computer Science and Engineering Department** staff for their timely support and suggestions.

We are also thankful to our parents for their moral support and constant guidance made our efforts fruitful. Last but not the least, our friends, without their constructive criticisms and suggestions we wouldn’t have been able to complete successfully.

**PROJECT ASSOCIATES**

**SHREYAS T M (18CS088)**

**VIVEK V R (18CS117)**

**THEJASWI K (18CS104)**

**VINOD S C (18CS114)**

**Abstract**

COVID-19 pandemic has affected the economy and changed the human way of life, disrupting everyone's mental, physical, and financial well-being. Many of the fastest-growing economies are strained owing to the severity and communicability of the epidemic. Because of the increasing diversity of cases and the resulting burden on healthcare practitioners and the government, therefore, predicting the number of infected COVID-19 cases which could be useful in planning the required hospital resources in the future. In this project, we focussed on information-led methods of estimating the numbers of COVID-19 confirmed cases in the country and their implications in the future, using Linear Regression model. Hence, this work would help the decision makers to understand the upcoming of the pandemic trajectory in the country and take necessary actions for the effect of interventions.

**CONTENTS :- Page Number**

**Chapter-1 Introduction 6-7**

[1](#_TOC_250001).1 Motivation  **7**

1.2 Advantages  **7**

**Chapter-2 Literature Review 8-10**

**Chapter-3 Requirements**

3.1 Project Requirements **10**

3.1 Hardware Requirements **10**

**Chapter-4 System Design 11-13**

4.1 Objectives **11**

4.2 Methodology/Planning and Working **11**

4.3 Modules **12**

4.3 Algorithm for prediction  **13**

4.4 Steps to implement Linear regression **14**

**Chapter-5 Output 15-19**

**Chapter-6 6.1 Conclusion 20**

**6.2 References 21**

**CHAPTER 1**

**1.INTRODUCTION:-**

At the end of December 2019 in Wuhan, China, it was first reported that a human infection was caused by a novel coronavirus (nCov) or Wuhan virus or 2019-nCov . One of the biggest challenges of this epidemic is the human-to-human transition of nCov. The coronavirus (COVID-19) infected cases increase at an exponential rate worldwide. On 30 January 2020, the World Health Organization (WHO) issued a worldwide health emergency warning notice , labeling that 2019-nCoV is of urgent global concern. The disease and mortality rates for the COVID-19 are uncertain at the early stage especially for young ones and aged people. WHO has estimated the reproduction factor of nCov is 2.7. In demand to control the extensive and quick spread of the nCov, public health sectors took reliable preventative measures and imposed curfew or lockdown infested cities in China, United States, India, and other countries also. This is to limit the social distance between people and to avoid the broadcast of this novel virus via humans to humans.

The best way to prevent and slow down transmission is maintaining social distance. We have to protect our self and others from infection by washing our hands or using sanitizers and avoid touching face. The number of COVID-19 cases in India are 67,161 and the death toll is 2,212 by 11th MAY 2020, as per the Worldometer data. Worldwide 4,180305 people have been attacked by virus and the total number of deaths caused by disease now are 283,865. There are very less number of COVID-19 test kits available in hospitals which are not at all sufficient for the increasing cases. Artificial Intelligence is actually dominant tool in the fight against the COVID-19 crisis. AI has subdomain like Machine Learning .It helps in diagnose and predict COVID-19. ML Techniques are useful in tracking COVID cases and predicting.The aim of this project is to develop a prediction model to predict the covid cases.

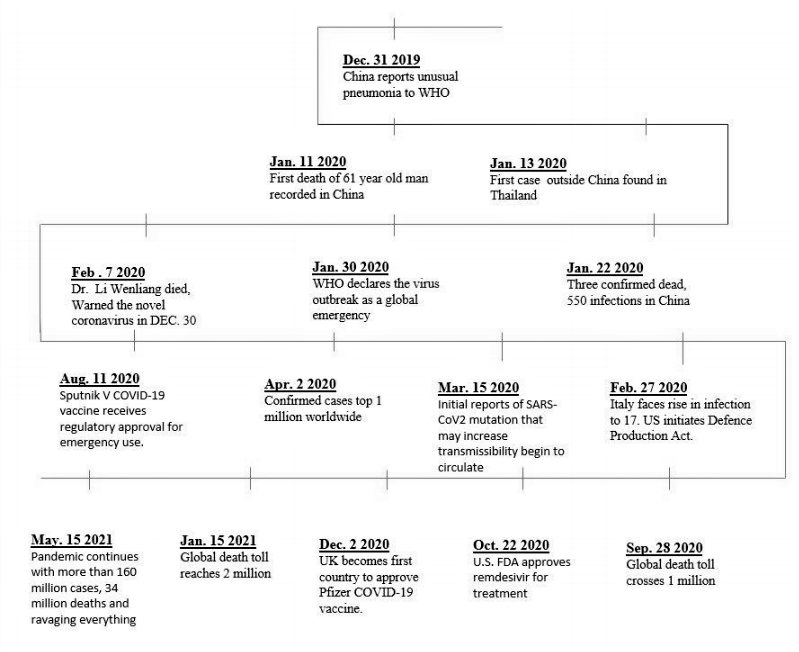


Figure : History of covid 19

**1.1 MOTIVATION:-**

The prediction of covid-19 cases is playing an important role during the ongoing COVID-19 pandemic. The first case of a COVID-19 affected patient was reported in Wuhan, China. Within one month of the outbreak, the number of positive cases and deceased rose at an exponential rate. As the pandemic situation has taken over the world, the main goal of this project is to develop a machine learning model that could predict number of COVID cases of a state in future. To develop such a model, a literature study alongside an experiment is set to identify a suitable algorithm. To assess the features that impact the prediction model.

**1.2 ADVANTAGES:-**

Following are the benefits or advantages of predicting the covid cases:

* Prediction of COVID-19 by using Machine Learning could help increase the speed of disease identification resulting in reduced mortality rate.
* Analyzing the results obtained from experiments, linear regression was identified to perform better compared to other algorithms.

**CHAPTER 2**

**2. LITERATURE SURVEY**:-

**A novel approach to predict COVID-19 using support vector machine by** [**Soham Guhathakurata**](https://www.ncbi.nlm.nih.gov/pubmed/?term=Guhathakurata%20S%5BAuthor%5D)**,** [**Souvik Kundu**](https://www.ncbi.nlm.nih.gov/pubmed/?term=Kundu%20S%5BAuthor%5D) **,** [**Arpita Chakraborty**](https://www.ncbi.nlm.nih.gov/pubmed/?term=Chakraborty%20A%5BAuthor%5D) **and**[**Jyoti Sekhar Banerjee**](https://www.ncbi.nlm.nih.gov/pubmed/?term=Banerjee%20JS%5BAuthor%5D)Published online 2021 May 21. **.**

The proposed method attempts to predict COVID-19 using a coined set of symptoms based on their severity and frequency. The key attributes in the feature set incline the prediction toward a particular class. The prediction has been classified into three classes, which are not infected, mildly infected, and severely infected. Among all the symptoms of COVID-19 cases, only a handful of them have been chosen. These includes are the most common symptoms (fever, breathing rate, cough) across 90% of the confirmed cases. We apply the support vector machine (SVM) classifier to classify the features/symptoms into the mentioned classes. In this project, we have also performed a comparative study on popular supervised learning models using visual programming. Based on the crucial impact of the symptoms, they have applied the support vector machine classifier to classify the patient's condition in no infection, mild infection, and serious infection categories and they have achieved an accuracy of 87% in predicting the cases.

**A novel approach to predict COVID-19 using KNN classification by Vamsidhar yendapalli and A Usha ruby.** Published online September 2020.

In this project they have used KNN classification algorithm to predict the COVID-19 confirmed cases. The k-fold cross-validation resampling technique is used to validate the prediction model. The prediction scores of each algorithm are evaluated with performance metrics such as prediction accuracy, precision, recall, mean square error, confusion matrix, and kappa score. For the given dataset, the k-nearest neighbour (KNN) classification algorithm produces 80.4 % of predication accuracy and 1.5 to 3.3 % of improved accuracy over other algorithms. The KNN algorithm predicts 92 % (true positive rate) of the deceased cases correctly with 0.077 % of misclassification. Further, the KNN algorithm produces the lowest error rate as 0.19 on the prediction of accurate COVID-19 cases than the other algorithm. Also, it produces the receiver operator characteristic curve with the output value of 82 %.

**A novel approach to predict COVID-19 using ANN based model Hamid Reza Niazkar and Majid Niazkar.** Published online 23 November 2020. Article number: 50 (2020).

The ANN-based models were utilized to estimate the confirmed cases of COVID-19 in India. These models exploit historical records of confirmed cases. The COVID-19 data were divided into a train part and a test part. The former was used to train the ANN models, while the latter was utilized to compare the purposes. The data analysis shows not only significant fluctuations in the daily confirmed cases but also different ranges of total confirmed cases observed in the time interval considered. Based on the obtained results, the ANN-based model that takes into account the previous 14 days outperforms the other ones. The accuracy of this model is 84-87%.

**Prediction using Linear regression**

Linear regression is a technique that have been used for this project. Linear regression has been used to predict the number of confirmed cases in a state of a country. By this prediction we can be getting a knowledge to how get protection and preventing the disease by all means.

In this project, a compartmental epidemiological model, named linear regression model, has been proposed for the prediction of COVID-19 peak in India and their states with the highest number of total cases. Further, short-term predictions have also been computed using this model, and the results are compared with the real world cases. The accuracy of this model is 89%.

**Chapter 3 POJECT REQUIREMENTS**

**3.1 SOFTWARE REQUIREMENTS**

* Python
* IDE(Jupyter notebook)
* Pandas
* Numpy
* Matplotlib
* Seaborn
* Sklearn
* Linear regression model

**Python**

Python is a high level and effective general use programming language. It supports multi-paradigms. Python has a large standard library which provide tools suited to perform various tasks. Python is a simple, less-clustered language with extensive features and libraries. Different programming abilities are utilized for performing the experiment in our work. In this thesis, the following python libraries were used.

* Pandas - It is a python package that provides expressive data structures designed to work with both relational and labelled data. It is an open source python library that allows reading and writing data between data structures.
* Numpy - It is an open source python package for scientific computing. Numpy also adds fast array processing capacities to python.
* Matplotlib - It is an open source python package used for making plots and 2D representations. It integrates with python to give effective and interactive plots for visualization.
* Sklearn - It is an open source python machine learning library designed to work alongside Numpy. It features various machine learning algorithms for classification, clustering and regression.
* Seaborn –It is a library in python predominantly used for making statistical graphics and is a data visualization library built on top of matplotlib and closely integarated with pandas data structures in python.

**3.2 HARDWARE REQUIREMENTS**

* Any Operating System with latest python 3.8.2 version.
* i5 or i7 processor
* RAM 4 or 8GB

**CHAPTER 4**

**SYSTEM DESIGN**:-

**4.1Objectives:-**

We aim to analyze data on the number of infected people in each Indian state (restricted to only those states with enough data for prediction) and predict the number of confirmed cases in that state. We hope that such state wise predictions would help the state governments better channelize their limited health care resources.

The main objectives are:

• Identifying the most suitable machine learning technique for prediction, to perform on clinical reports of patients.

• Preparing a machine learning model that could make accurate predictions of COVID-19 in patients.

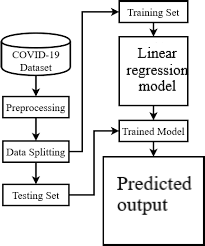
• Identifying the features that affects the prediction of COVID-19 in patients.

**4.2 Methodology/Planning and Working**

Data was collected in csv file and uploaded in jupyter notebook and analysed with the python 3.8.2 software.

Input : Total number of confirmed cases, death cases, cured cases.

Output : Predicted confirmed cases of a state.



**Figure :** Data flow diagram for proposed algorithm

**4.3 Functions**

* **df.head()**

This function returns the first n rows for the object based on position. It is useful for quickly testing if your object has the right type of data in it.

* **df.columns = [' ']**

The column labels of the DataFrame.

* **df.tail()**

This function returns the last n rows for the object based on position. It is useful for quickly testing if your object has the right type of data in it.

* **plt.show()**

The show() function in pyplot module of matplotlib library is used to display all figures.

* **sns.lineplot( )**

Draw a line plot with the possibility of several semantic groupings. The relationship between x and y can be shown for different subsets of the data using the hue, size, and style parameters.

* **sns.barplot( )**

A barplot is basically used to aggregate the categorical data according to some methods and by default it’s the mean. It can also be understood as a visualization of the group by action.

* **x\_train,x\_test,y\_train,y\_test=train\_test\_split(x,y, )**

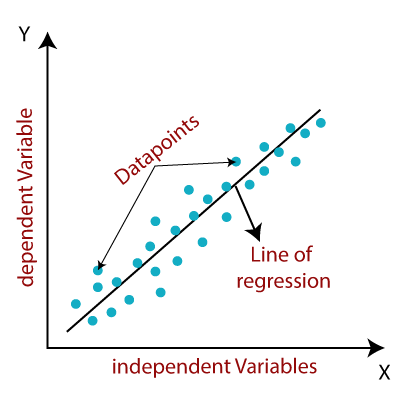
The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model.

* **lr.fit(np.array(x\_train).reshape(-1,1),np.array(y\_train).reshape(-1,1))**

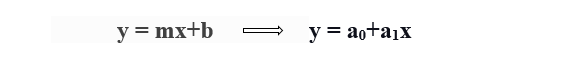
Fitting a linear regression means finding the coefficients and bias that best model the relationship between x and y.

**4.4 Algorithm for prediction**

Linear regression is a quiet and simple statistical regression method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression. If there is a single input variable (x), such linear regression is called **simple linear regression**. And if there is more than one input variable, such linear regression is called **multiple linear regression.** The linear regression model gives a sloped straight line describing the relationship within the variables.



The above graph presents the linear relationship between the dependent variable and independent variables. When the value of x (independent variable) increases, the value of y (dependent variable) is likewise increasing. The red line is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best.

To calculate best-fit line linear regression uses a traditional slope-intercept form.y= Dependent Variable.  
x= Independent Variable.  
a0= intercept of the line.  
a1 = Linear regression coefficient.

**Benefits** of using linear regression

* widely used
* runs fast
* easy to use (not a lot of tuning required)
* highly interpretable
* basis for many other methods

### 4.5 STEPS TO IMPLEMENT LINEAR REGRESSION

### Import some required libraries.

### Define the dataset.

### Plot the data points.

### Plot the regression line.

### Predicting the values.

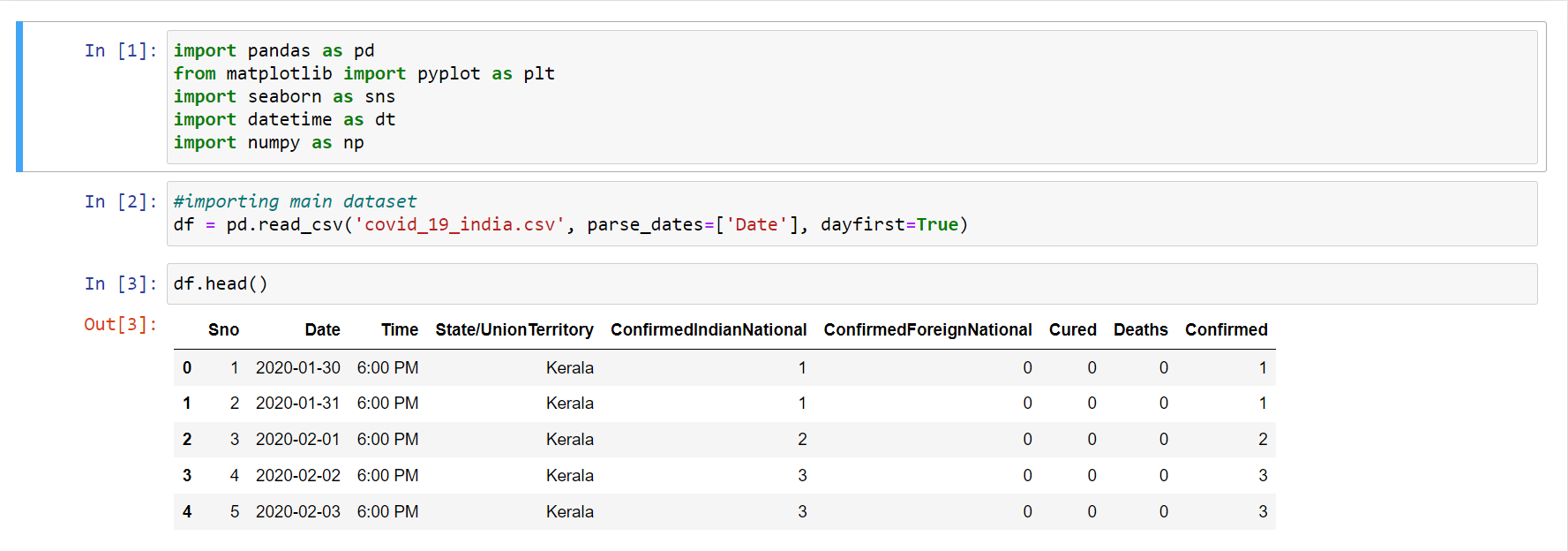
### 

### Figure : Implementation of Linear regression

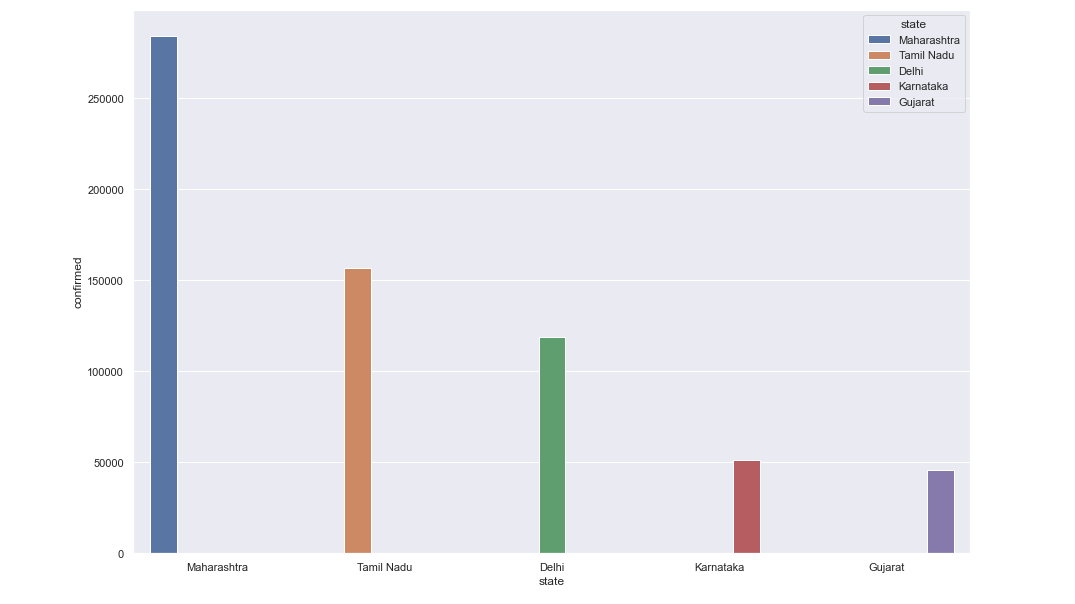
**CHAPTER 5**

**OUTPUT**

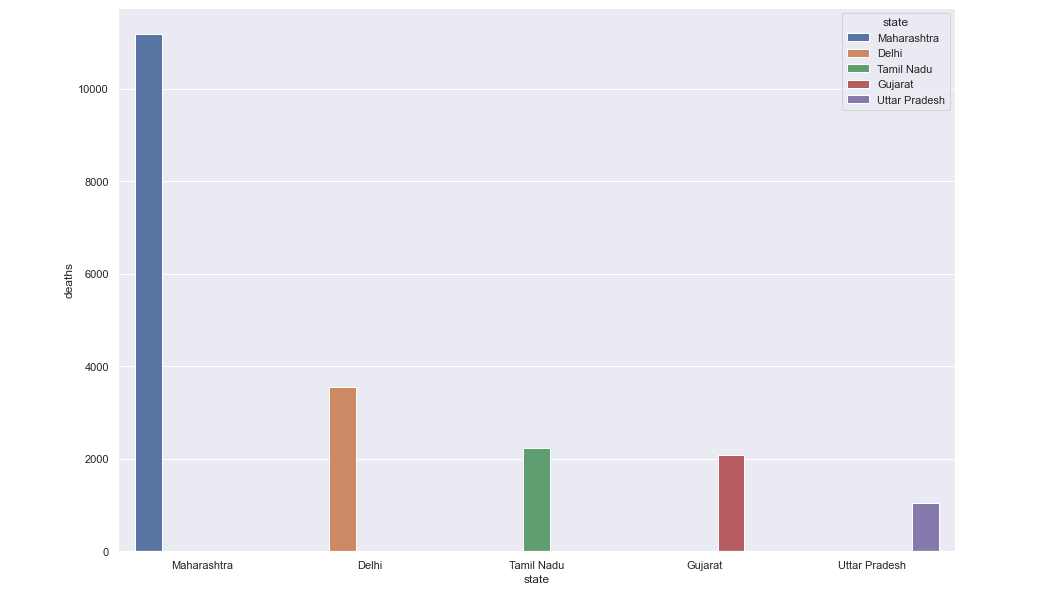
**[1] Importing libraries and reading dataset**



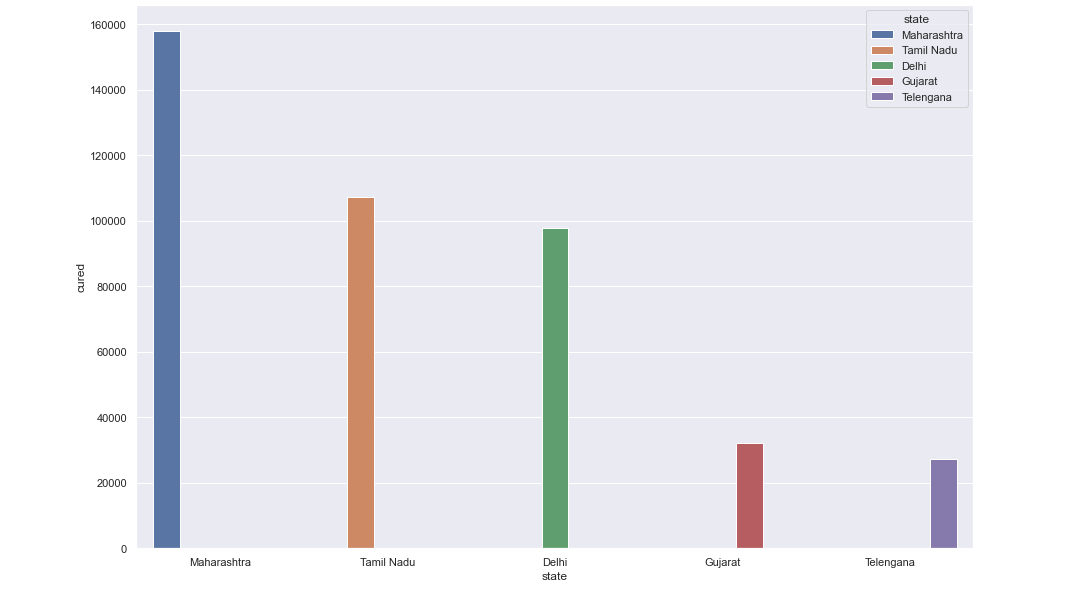
**[2]** **Plotting a bar graph for top 5 states with maximum number of confirmed cases.**

****

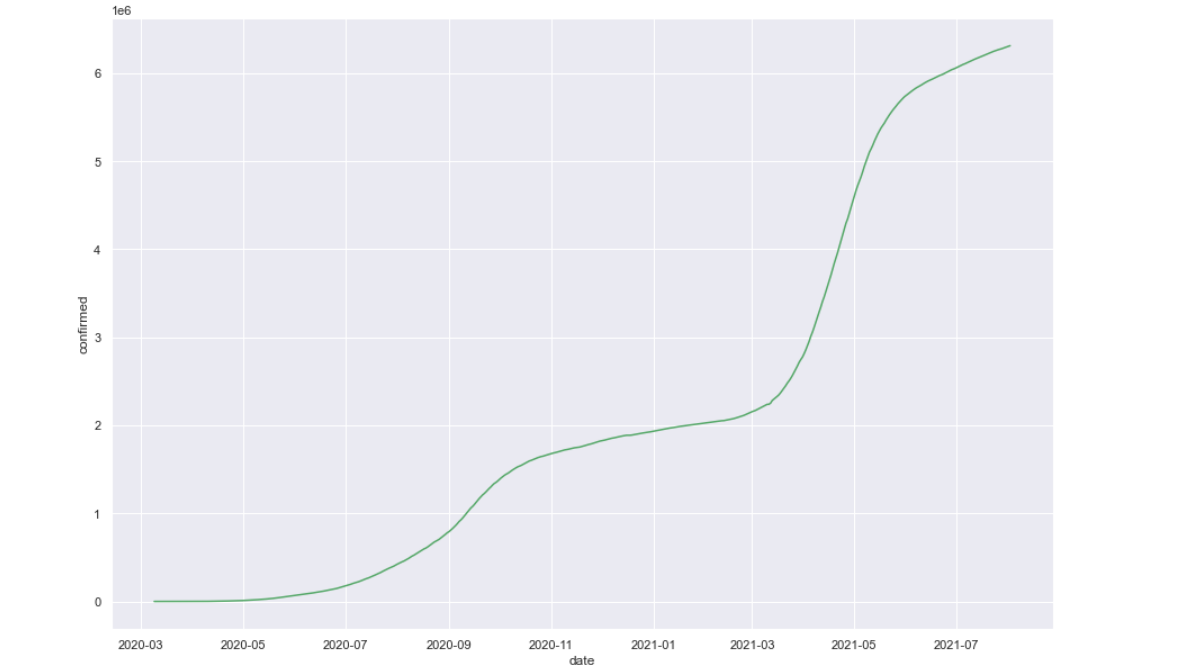
**[3]** **Plotting a bar graph for top 5 states with maximum number of death cases.**

****

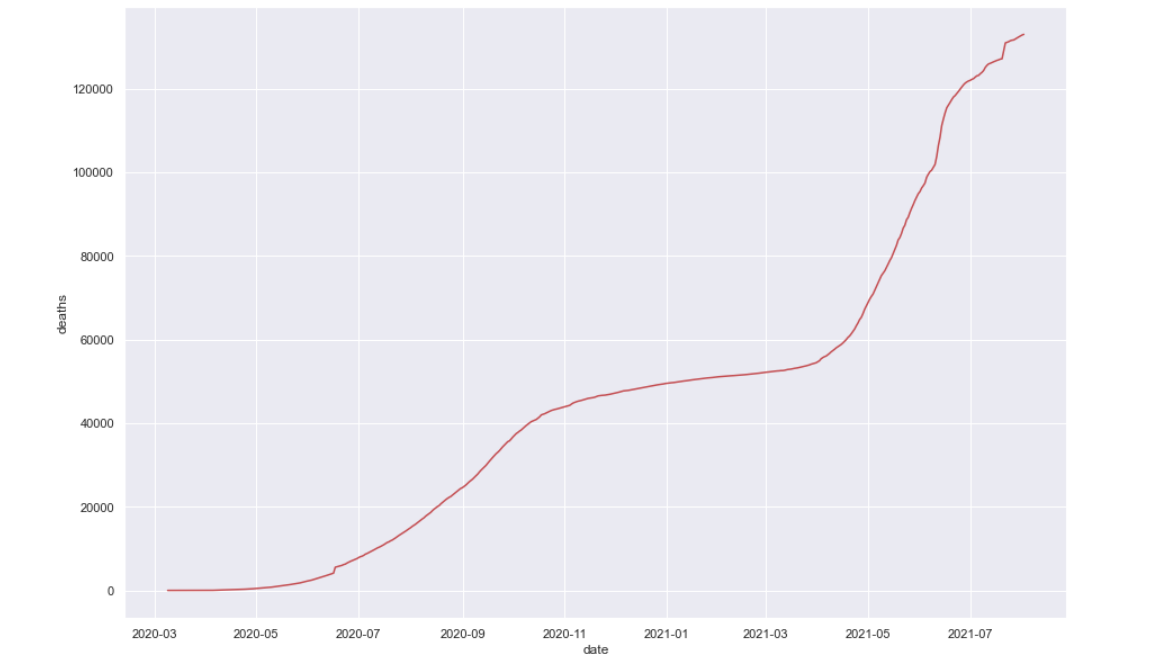
**[4]** **Plotting a bar graph for top 5 states with maximum number of cured cases.**

****

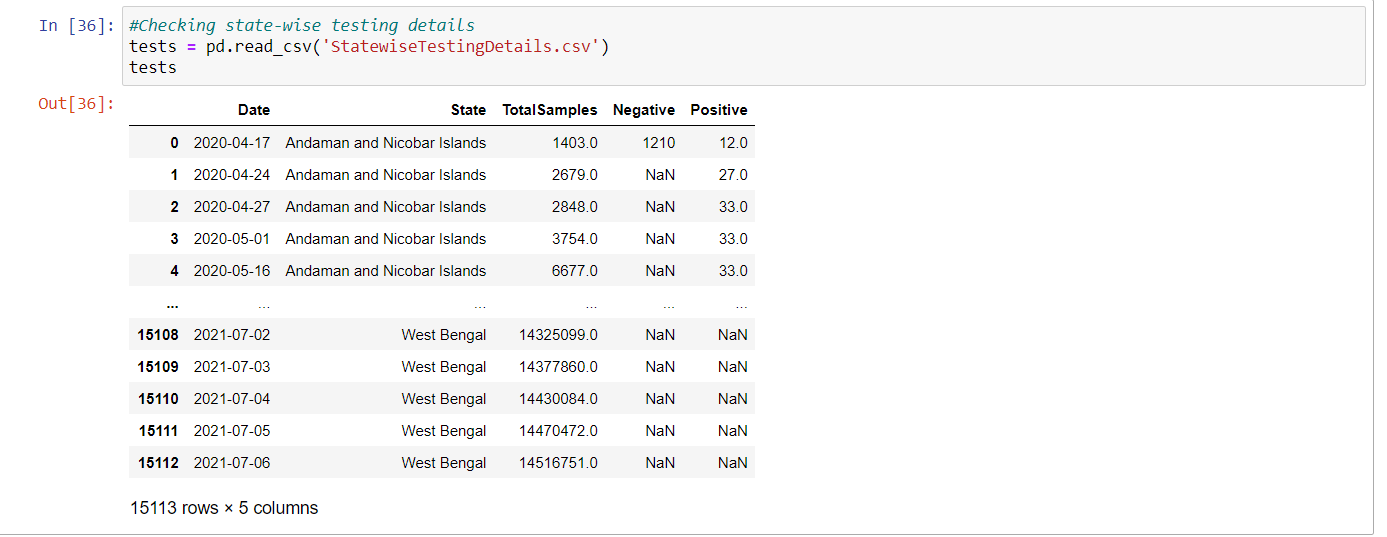
**[5]** **Plotting a line graph for confirmed cases in Maharashtra.**

****

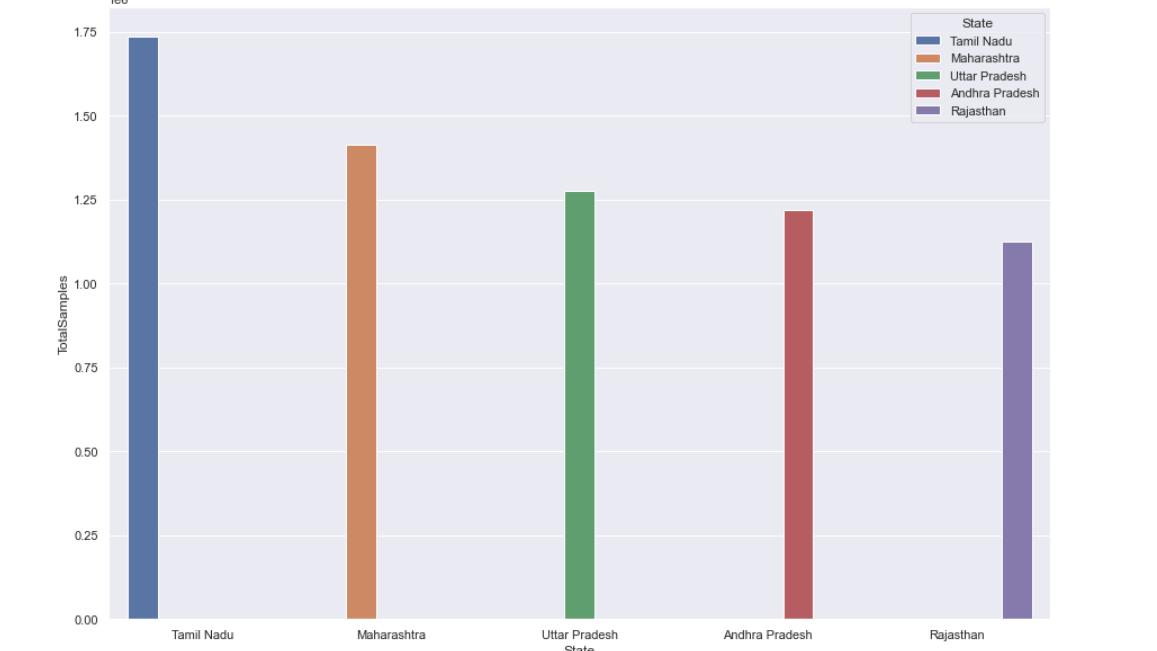
**[6]** **Plotting a line graph for death cases in Maharashtra.**

****

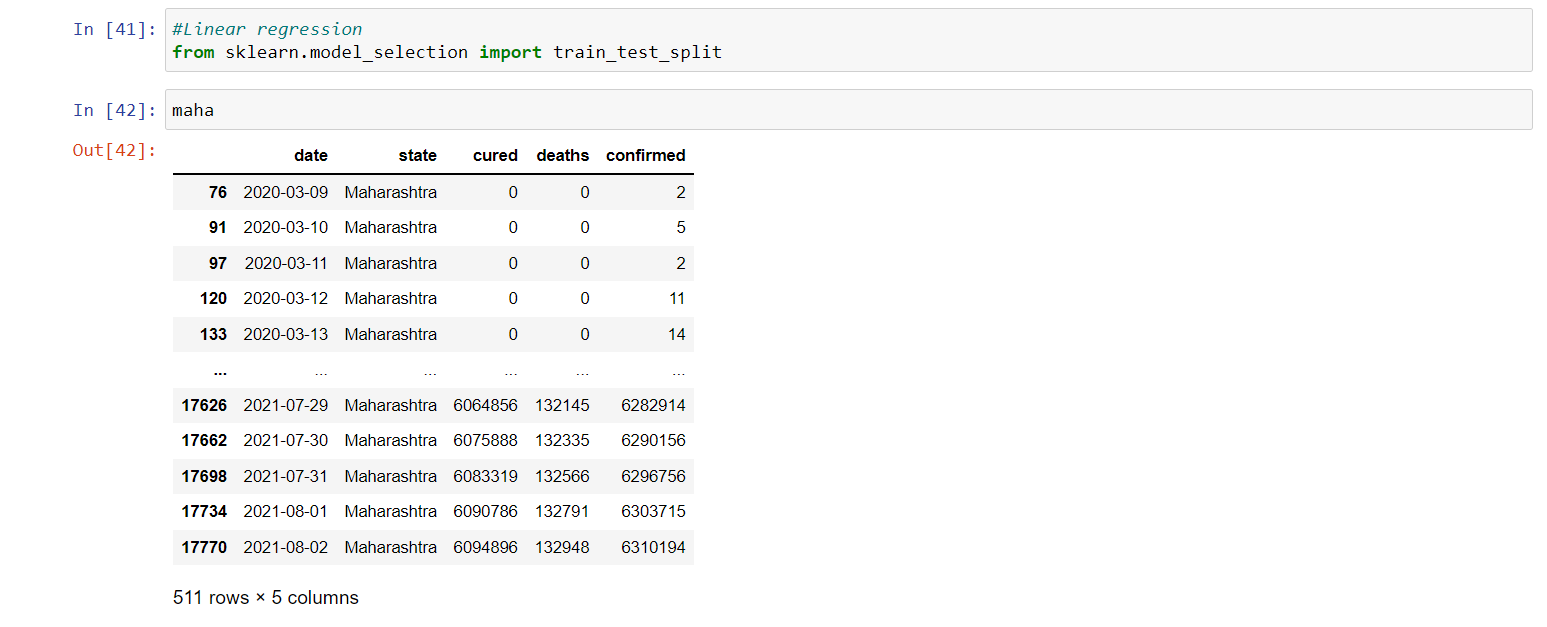
**[7]** **Reading the dataset of state-wise testing details of all the states in India**.



**[8]** **Plotting a bar graph for states with maximum number of test cases.**



**[9]** **Implementing linear regression using sklearn model.**

****

**[10]** **Predicting the confirmed cases using linear regression.**

****

**CHAPTER 6**

**6.1 CONCLUSION**

This study investigates how new COVID-19 cases can be predicted while considering the historical data of COVID-19 cases alongside the external factors that affect the spread of the virus. The effectiveness and superiority of the developed algorithm (Linear regression) are demonstrated by conducting experiments using data collected for top five affected states in India. The results show an improved accuracy if compared with the existing methods. Moreover, the experiments are extended to make future prediction of the affected COVID-19 cases. The predicted COVID-19 cases help in providing some recommendations for both the government and people of the affected countries. This project provides a novel way for predicting the number of COVID-19 cases. However, there are some venues that might be suitable for future directions. For example, predicting the number of deaths could be one direction. Another direction might be predicting the number of recovered people.

**6.2 REFERENCES**

1. [**https://www.kaggle.com/sudalairajkumar/covid19-in-india**](https://www.kaggle.com/sudalairajkumar/covid19-in-india)
2. [**https://www.greatlearning.in/academy/learn-for-free/courses/predicting-covid-19-in-india-using-machine-learning**](https://www.greatlearning.in/academy/learn-for-free/courses/predicting-covid-19-in-india-using-machine-learning)
3. [**https://www.tutorialspoint.com/machine\_learning\_with\_python/regression\_algorithms\_linear\_regression.htm**](https://www.tutorialspoint.com/machine_learning_with_python/regression_algorithms_linear_regression.htm)
4. **reportCoronavirus disease (COVID-2019) situation reports. World Health Organization.Available at::**[**https://www.who./emergencies/diseases/novel-coronavirus-2019/situation-reports/**](https://www.who./emergencies/diseases/novel-coronavirus-2019/situation-reports/)**[Accessed on: 26nd March 2020].**
5. ***INDIA COVID-19 TRACKER.* 2020.**[**https://www.covidindia.org/**](https://www.covidindia.org/)**[Online] Available at: Accessed on: 26nd March 2020. [**[**Google Scholar**](https://scholar.google.com/scholar_lookup?title=INDIA+COVID-19+TRACKER&publication_year=2020&)**]**
6. [**https://ghrp.biomedcentral.com/articles/10.1186/s41256-020-00175-y#Abs1**](https://ghrp.biomedcentral.com/articles/10.1186/s41256-020-00175-y#Abs1)